Claims

We claim:

1. In a computer data processing system, a method for clustering data in a database comprising the steps of:

a) reading data records having both discrete and ordered attributes from a database storage medium and bringing a portion of the data records into a rapid access memory;

b) initializing a cluster model that characterizes the data within the database wherein the cluster model includes a table of probabilities for the enumerated or discrete data attributes of the data records for each cluster of a multiple number of clusters that make up the cluster model and wherein the cluster model for data attributes that are ordered comprises a mean and covariance for each cluster;

c) updating the cluster model from the database records brought into the rapid access memory;

d) summarizing at least some of the database records in the rapid access memory and storing a summarization within the rapid access memory;

e) evaluating a criteria to determine if further data should be accessed from the database to further cluster data records in the database; and

f) based on the evaluating step reading an additional number of records from the database storage medium and bringing said additional number of records into the rapid access memory for further updating of the cluster model.

2. The method of claim 1 wherein the step of updating the cluster model includes the step of adjusting the table of discrete attribute probabilities for a cluster by

calculating a weighted sum of the data records brought into the rapid access memory and a weighted sum for data records already summarized in the cluster model.

3. The method of claim 1 wherein the step of updating the cluster model includes the step of adjusting a data structure of ordered attribute mean and covariance values by calculating a weighted sum of the mean and covariance values of database records brought into the rapid access memory and the mean and covariance values for records already summarized in the cluster model.

4. The method of claim 1 wherein the step of updating the cluster model includes adjusting the ordered attribute mean and spread values and the table of discrete attribute probabilities for a cluster by calculating a weighted sum of the mean and covariance values and probability values of database records brought into the rapid access memory and the mean and covariance values and probability values for records already summarized in the cluster model.

5. The method of claim 1 wherein both the ordered and the discrete attributes are assigned a confidence interval and wherein the summarizing step summarizes certain data records based upon the confidence interval.

6. The method of claim 5 wherein the step of summarizing the database records includes the step of determining whether a data point is suitable for summarization by performing a perturbation of the cluster model probabilities and verifying that the data point is sufficiently described by the perturbed cluster model probabilities.

7. The method of claim 5 wherein the step of summarizing the database records includes the step of determining whether a data point is suitable for summarization

by comparing the probability that a data point belongs to a cluster with a threshold probability value.

8. The method of claim 5 wherein the step of summarizing the data base records includes the step of performing a non-scalable clustering method on the data points remaining in rapid access memory after some of the database records have been summarized and storing the results of the non-scalable clustering method in rapid access memory.

9. The method of claim 5 further including the step of constructing a model of the database based on the cluster probability tables, the summarizations of data points in rapid access memory, and data points in rapid access memory which have been neither compressed or summarized.

10. The method of claim 9 wherein the step of constructing a model of the database comprises the steps of:

a) resetting a new model data structure to zero;

b) determining a weighted contribution to the new model for each unsummarized data point; and

c) determining a weighted contribution of the summarized data points to the new model.

11. The method of claim 10 further comprising the steps of:

a) providing an old model based upon a past modeling iteration;

b) comparing the old model to the new model; and

c) terminating the modeling process when the old model and new model are sufficiently similar.

12. The method of claim 1 wherein a probability that a data record belongs in a cluster for data records extracted from the database is calculated using a covariance matrix for the continuous attributes of the record.

13. The method of claim 1 wherein each cluster in the cluster model is characterized by a datapoint number for the cluster, a mean for each ordered data attribute, a covariance for each ordered data attribute and a probability table for each discrete data attribute and further wherein each data record read from the database storage medium contributes to an updating of the cluster model for at least one cluster.

14. The method of claim 13 wherein there are K clusters in the cluster model and wherein each data record contributes to a cluster model for each of the K clusters.

15. The method of claim 1 wherein the step of accessing database records is performed using a sequential scan of the database.

16. The method of claim 1 wherein the step of accessing database records is performed using a random index generator that does not repeat.

17. In a computer data mining system, apparatus for evaluating data in a database comprising:

a) one or more data storage devices for storing a database of data records on a storage medium; said data records including attributes of both discrete or enumerated data and ordered data;

b) a computer having a rapid access memory and an interface to the storage devices for reading data from the storage medium and bringing the data into said rapid access memory for subsequent evaluation; and

c) said computer comprising a processing unit for evaluating at least some of the data records in the database and for characterizing the data records into multiple numbers of data clusters; said processing unit programmed to retrieve a subset of data from the database into the rapid access memory, evaluate the subset of data to further characterize the database clustering using a clustering criteria, and produce a summarization of at least some of the retrieved data records before retrieving additional data records from the database; said computer producing a cluster model that includes cluster probabilities for the discrete attributes and cluster means and covariance information for the ordered data in the rapid access memory during data clustering.

18. The apparatus of claim 17 wherein said processing unit updates said cluster model criteria based on said subset of said data records and previously summarized data from the database.

19. The apparatus of claim 17 wherein said processing unit is further programmed to summarize certain data and to summarize certain other data according to subclusters having means, covariances, and probabilitity tables characterizing each of said subclusters.

20. The apparatus of claim 17 further comprising an output means for outputting a model of said database created by said characterization of data into clusters.

21. A computer readable medium having stored thereon a data structure, comprising:

a) a first data portion containing a model representation of data records stored in a database, wherein at least some of the database records include mixed data that includes both discrete data fields and continuous data fields;

b) a second data portion containing sufficient statistics of a portion of the data records in the database; and

c) a third data portion containing individual data records obtained from the database for use with the sufficient statistics to determine said model representation contained in the first data portion.

22. The data structure of claim 21 wherein said model representation comprises a set of clusters to which data records may be assigned based on the degree to which each cluster describes the data record.

23. The data structure of claim 22 wherein each of said clusters is represented by a datapoint number, a mean for each ordered attribute, a spread for each ordered attribute, and a probability table for each discrete attribute.

24. The data structure of claim 22 wherein each of said data records may be assigned to only one of said clusters.

25. The data structure of claim 21 wherein a second data portion containing sufficient statistics of a portion of the data records in the database is organized by cluster and includes a data record number, a mean for each ordered attribute, a spread for each ordered attribute, and a probability table for each discrete attribute of those records summarized in said sufficient statistics.

26. The data structure of claim 22 wherein each of said data records may be assigned to each of said clusters with a probability based on the degree to which a given cluster describes said data record.

27. The data structure of claim 21 wherein said sufficient statistics comprises a set of clusters to which data records may be assigned based on the degree to which each cluster describes the data record.

28. The data structure of claim 27 wherein the sufficient statistics for each of said clusters is represented by a datapoint number, a mean for each ordered attribute, a spread for each ordered attribute, and a probability table for each discrete attribute which in combination with individual data records is used to produce the cluster model representation.

29. The data structure of claim 28 wherein each of said data records may be summarized with a data summarization associated with a data cluster or may also be assigned to a data summarization associated with a subcluster or may be left as a vector data record.

30. A computer-readable medium having computer-executable components comprising:

a) a database component for interfacing with a database that stores data records containing both enumerated or discrete and ordered values;

b) a rapid access memory component for storing at least of subset of said data records gathered from the database for processing;

c) a modeling component for constructing and storing a model of said database by determining if a data record is sufficiently described by any of several clusters using cluster criteria and for updating said cluster model based on said data records and for evaluating whether further of said data records should be moved from said database into said rapid access memory for modeling.

31. The computer readable medium of claim 30 wherein said database component is adapted to store and said modeling component to construct a model of data records containing both enumerated or discrete and ordered values.

32. The computer readable medium of claim 30 wherein said modeling component is adapted to store said model of said database in the form of a datapoint number, a table of means, a table of spreads, and a table of probabilities for each cluster of said cluster model.

33. The computer readable medium of claim 30 wherein said modeling component is adapted to compare a new model to a previously constructed model to evaluate whether further of said data records should be moved from said database into said rapid access memory for modeling.

34. The computer readable medium of claim 30 wherein said modeling component is adapted to update said cluster model by calculating a weighted contribution by each of said data records in said rapid access memory.